|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | **discrete data** |
| Results of rolling a dice | **discrete data** |
| Weight of a person | **continues data** |
| Weight of Gold | **continues data** |
| Distance between two places | **continues data** |
| Length of a leaf | **continues data** |
| Dog's weight | **continues data** |
| Blue Color | **Categorical data** |
| Number of kids | **discrete data** |
| Number of tickets in Indian railways | **discrete data** |
| Number of times married | **discrete data** |
| Gender (Male or Female) | **Categorical data** |

Q1) Identify the Data type for the Following:

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

|  |  |
| --- | --- |
| Data | Data Type |
| Gender | **Nominal** |
| High School Class Ranking | **Ordinal** |
| Celsius Temperature | **Interval** |
| Weight | **Ratio** |
| Hair Color | **Ratio** |
| Socioeconomic Status | **Ordinal** |
| Fahrenheit Temperature | **Ratio** |
| Height | **Ratio** |
| Type of living accommodation | **Interval** |
| Level of Agreement | **Ordinal** |
| IQ(Intelligence Scale) | **Interval** |
| Sales Figures | **Interval** |
| Blood Group | **Ratio** |
| Time Of Day | **Ordinal** |
| Time on a Clock with Hands | **Interval** |
| Number of Children | **Ratio** |
| Religious Preference | **Ordinal** |
| Barometer Pressure | **Interval** |
| SAT Scores | **Interval** |
| Years of Education | **Ratio** |

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

**Ans: Total possible outcomes while tossing three coins are; {HHH, HHT, HTH, TTH, THT, THH, HTT, TTT}.**

**No. of possible outcomes are=8, now we want to know about the probability of getting two heads and one tail while tossing, Let A= event of getting two heads and one tail. i.e., A= {HHT, HTH, THH}**

**Therefore P(A)= No. of favorable events\No. of total events**

**P(A)=3\8=37.5%**

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2 and 3

**Ans: Total no. possible outcomes while rolling two dice=36**

1. **Let E= event of having sum equal to 1**

**There are no possible outcomes having sum equal to 1. Because while rolling two dices the minimum sum value is 2 for (1,1). So, there is no chance of getting 1. that means probability of sum equal to 1= P(E)=0.**

1. **Let E= Sum is less than or equal to 4.**

**Favorable outcomes for sum are less than or equal to 4; {(1,1), (1,2), (1,3), (2,1), (3,1), (2,2)}. Therefore, probability of sum is less than or equal to zero =P(E)=6\36=16.66%**

1. **Let E= Event of sum is divisible by 2 and 3. Favorable outcomes for sum is divisible by 2 and 3 are; {(1,5), (5,1), (2,4), (4,2), (3,3), (6,6)}. Therefore, probability of sum is divisible by 2 and 3 = P(E)=6\36=16.66%**

Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

**Ans: Total no. of balls=7. No. of ways drawing 2 balls out of 7 balls =7C2=21.**

**Let E= Event of drawing 2 balls none of which is blue.**

**Here are 2 blue balls so we don’t need to consider that 2 balls. Therefore no. of ways drawing 2 balls out of 5 balls=5C2=10. P(E)=10\21= 47.62%**

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

|  |  |  |
| --- | --- | --- |
| CHILD | Candies count | Probability |
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

**Ans: Expected no. of candies for a randomly selected child**

**E(X)=**

**= 1 × 0.015 + 4× 0.20 + 3 × 0.65 + 5× 0.005 + 6 × 0.01 + 2 × 0.12**

**= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24**

**= 3.090**

**= 3.09**

**Therefore expected no. of candies for a randomly selected child is 3.09**.

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points,Score,Weigh>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

**Ans:**

**Import pandas as pd**

**df=pd.read\_csv(“Q7.csv”)**

**df.describe()**

**df.mode()**

**df.var()**

**df.std()**

**range=df[["Points","Score","Weigh"]].max()-df[["Points","Score","Weigh"]].min()**

**range**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weigh** |
| **Mean** | **3.596563** | **3.217250** | **17.848750** |
| **Median** | **3.695000** | **3.325000** | **17.710000** |
| **Mode** | **3.92** | **3.44** | **17.02** |
| **Variance(s2)** | **0.285881** | **0.957379** | **3.193166** |
| **Standard Deviation(s)** | **0.534679** | **0.978457** | **1.786943** |
| **Range** | **2.17** | **3.911** | **8.4** |

**Inference: The data is based upon different types of cars with its points, score and weigh. Average in the points of the cars is around 3.59, average score is 3.21 and the average weigh of the listed cars are around 17.84.**

**Here the median or the centric value of the points in the cars data is 3.69 that means 3.69 equally splits the points value into two equally sized groups. Similarly, we can say that median of the score of the data set is around 3.32 and the median of the weigh of the data set is around 17.710.**

**Mode is nothing but mostly occurred value in the data set. So, the mode of the cars dataset for points is 3.92, mode of the score is 3.44 and mode of the weigh in the dataset is 17.02.**

**Standard deviation of a dataset speaks about how well the datapoints got deviated from the mean value so the deviation for points is around .534, for score is around .9784 and for weigh is around 1. 786, So comparing with point and score, the weigh of the cars data is more deviated from its mean value.**

Q8) Calculate Expected Value for the problem below

1. The weights (X) of patients at a clinic (in pounds), are

108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

**Ans: Total no. of patience=9.Here we want to know about the probability of each. Here we are chosen one of the patience among this at random. So, probability of getting any of them is 1\9.**

**Expected value of the weight of the patient**

**= (1/9) (108) + (1/9)110 + (1/9)123 + (1/9)134 + (1/9)135 + (1/9)145 + (1/9(167) + (1/9)187 + (1/9)199**

**= (1/9) (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)**

**= (1/9) (1308)**

**= 145.33**

**Therefore, expected value of the weight of the patient is 145.33.**

Q9) Calculate Skewness, Kurtosis & draw inferences on the following data

Cars speed and distance

Use Q9\_a.csv

**Ans: import pandas as pd**

**data=pd.read\_csv(“Q9\_a.csv”)**

**data.kurt(axis=0,skipna=True)**

**data.skew(axis=0,skipna=True)**

**Calculated values:**

**skewness of speed = -0.117510**

**skewness of distance= 0.806895**

**kurtosis of speed=-0.508994**

**kurtosis of distance=0.405053**

**Inference: The skewness of the speed is negative it means that the tail on the left side of the distribution is longer, most of the points are in right side and the mean value is less than the median. By going through the kurtosis of the speed, it is also negative means that the data has light tail and broad peak.**

**The skewness of the distance is positive it means that the tail on the right side of the distribution is longer, most of the points are in left side and the mean value is greater than the median. By going through the kurtosis of the distance, it is also positive means that the data has heavy tail and pointed peak.**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans: import pandas as pd**

**df1=pd.read\_csv(“Q9\_b.csv”)**

**df1.kurt(axis=0,skipna=True)**

**df1.skew(axis=0,skipna=True)**

**Calculated values:**

**skewness of SP = 1.611450**

**skewness of WT= -0.614753**

**kurtosis of SP= 2.977329**

**kurtosis of WT=0.950291**

**Inference: The skewness of the SP is positive it means that the tail on the right side of the distribution is longer, most of the points are in left side and the mean value is greater than the median. By going through the kurtosis of the SP, it is also positive means that the data has heavy tail and pointed peak.**

**WT is negatively skewed and the kurtosis is positive, which says that the distribution is more towards right. And the mean value is less than the median. Positive kurtosis means that the data has heavy tail and pointed peak.**

**Q10) Draw inferences about the following boxplot & histogram**



**Ans: The histogram speaks about the positive skewness of the distribution. Most of the values are clustered around the left tail of the distribution while the right tail of the distribution is longer. So, by regarding of the observation comparatively most of the points are in the range of 50 to 100 in chickWeight$weight (based on x- axis).**



**Ans: Here the box plot speaks about the positively skewness of the distribution because the median is closer to the lower quartile and also there are some outliers in the upper extreme.**

**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

**Ans:**

**population size=3,000,000 sample size(n)=2,000**

**Sample mean=200 sample standard deviation(s)=30**

**For calculating the t value:**

**From scipy import stats**

**From math import sqrt**

**stats.t.ppf(CI,df)**

**t value for 94%= 1.8818614764780113**

**t value for 96%=2.055089962825778**

**t value for 98%= 2.328214776106972**

**For calculating the confidence interval:**

**Stats.norm.interval(x,mean,scale=())**

**Confidence interval of 94% is [198.738325292158, 201.261674707842]**

**Confidence interval of 98% is [198.62230334813333, 201.37769665186667]**

**Confidence interval of 96% is [198.43943840429978, 201.56056159570022]**

**Q12)** Below are the scores obtained by a student in tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find mean, median, variance, standard deviation.
2. What can we say about the student marks?

**Ans: 1) Mean=40.705882**

**Median=40.00**

**Standard deviation=5.405607**

**Variance= 29.220588**

**2) Average score of the student is 40.71. That means most of the time he scores 40.71.**

Q13) What is the nature of skewness when mean, median of data are equal? 0 skewness and the distribution is symmetric

Q14) What is the nature of skewness when mean > median ? Positively skewed and the distribution is more towards the left

Q15) What is the nature of skewness when median > mean? Negatively skewed and the distribution is more towards the right

Q16) What does positive kurtosis value indicates for a data ? Positive values of kurtosis indicate that a distribution is peaked and possess thick tails.

Q17) What does negative kurtosis value indicates for a data? Negative values of kurtosis indicate that a distribution is flat and has thin tails

Q18) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

What is nature of skewness of the data?

What will be the IQR of the data (approximately)?   
  
**Ans: Most of the observations are on the high end of the scale, so the distribution is** [**skewed left**](https://stattrek.com/statistics/dictionary.aspx?definition=skewness)**. The** [**interquartile range**](https://stattrek.com/statistics/dictionary.aspx?definition=Interquartile%20range) **is indicated by the length of the box, which is 18 minus 10 or 8. And the median is indicated by the vertical line running through the middle of the box, which is roughly centred over 15. So the median is about 15.**

Q19) Comment on the below Boxplot visualizations?



Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

**Ans: Both the boxplots show some symmetry means that they are normally distributed. There is no outlier, median is around 262.5 something.**

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

**Ans: import pandas as pd**

**From scipy import stats**

**data=pd.read\_csv(“Cars.csv”)**

**To calculate mean and standard deviation**

**data.describe()**

**For P(MPG>38):**

**1-stats.norm.cdf(38,loc=34.422076,scale=9.131445)= 0.34759394041453007**

**For P(MPG<40):**

**stats.norm.cdf(40,loc=34.422076,scale=9.131445)= 0.7293498604157946**

**For P (20<MPG<50):**

**stats.norm.cdf(50,loc=34.422076,scale=9.131445)- stats.norm.cdf(20,loc=34.422076,scale=9.131445)= 0.8988689076273199**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

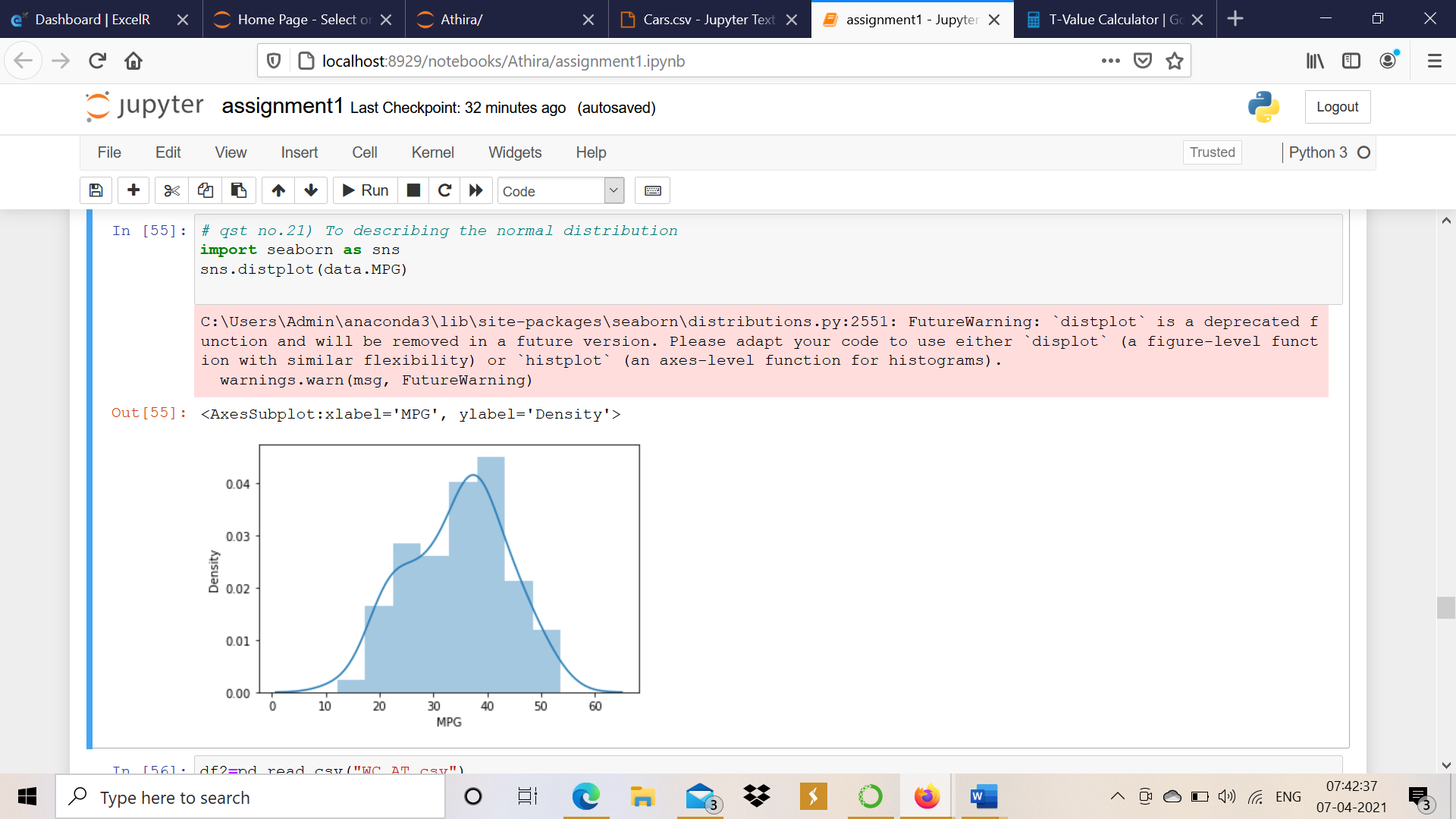
Dataset: Cars.csv

**Ans: import pandas as pd**

**import seaborn as sns**

**data=pd.read\_csv(“Cars.csv”)**

**sns.distplt(data.MPG)**



Here by using distplot of MPG we can see that MPG of cars shows some normal distribution.

1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

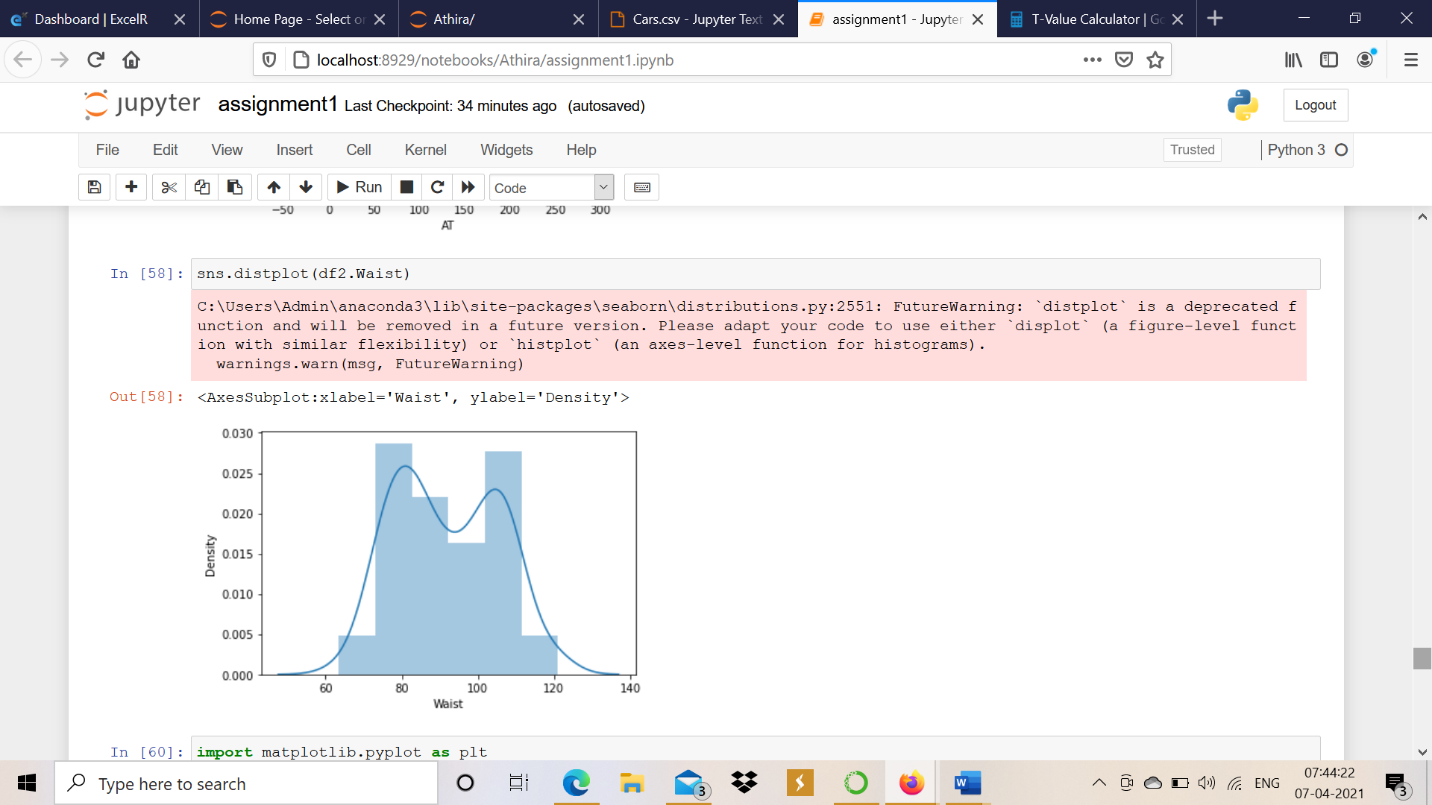
**Ans: import pandas as pd**

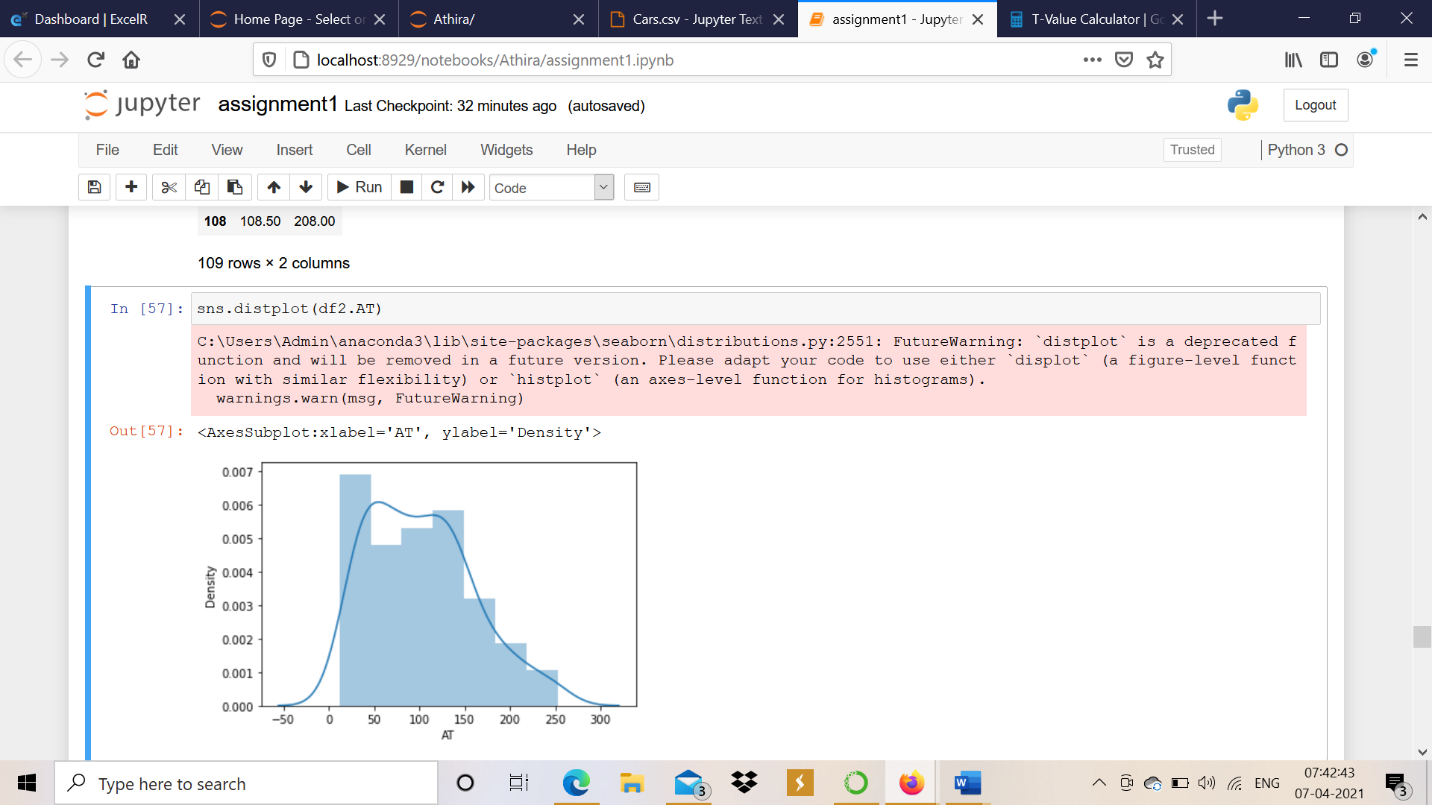
**import seaborn as sns**

**data=pd.read\_csv(“wc-at.csv”)**

**sns.distplt(data.Waist)**

**sns.distplot(data.AT)**





**Here by using distplots of Waist and AT we can see that both of them show some normal distribution.**

Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

**Ans: from scipy imports stats**

**Stats.norm.ppf(.95)**

**Z score for 90% confidence interval: 1.6448~1.645**

**from scipy imports stats**

**Stats.norm.ppf(.97)**

**Z score for 94% confidence interval:1.8807~1.881**

**from scipy imports stats**

**Stats.norm.ppf(.80)**

**Z score for 60% confidence interval: .84162~.842**

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

**Ans: from scipy imports stats**

**Stats.t.ppf(.975,df=24)**

**t score for 95% confidence interval:2.06389~2.064**

**from scipy imports stats**

**Stats.t.ppf(.98,df=24)**

**t score for 96% confidence interval:2.1715~2.172**

**from scipy imports stats**

**Stats.t.ppf(.995,df=24)**

**t score for 99% confidence interval:2.7969~2.797**

Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

**Ans: population mean(µ)=270**

**Sample mean (x̅) =260**

**Sample size(n)=18**

**Sample standard deviation(s)=90**

**t score = (x̅- µ)/(s/sqrt(n))**

**=(260-270)/(90/sqrt(18))**

**=-.471**

**From scipy import stats**

**Stats.t.ppf(-.471,17)**

**Probability of 18 randomly selected bulbs=.3498**